In [ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
pd.options.display.max_columns = None
```

LA GRAN PANDEMIA

ABSTRACT

Durante decadas y hasta siglos la humanidad ha estado luchando contra enfermedades infecciones capaces de llegar a ser pandemias o epidemias. Miles tratan de entender como es su funcionamiento y como se comportan en una poblacion. El desafio para la lucha contra estas enfermedades es saber de ante mano el comportamiento que va a tener la misma en la poblacion, su ciclo de vida, su capacidad de contagio, su capacidad de propagacion, su capacidad de mutacion, etc. Tambien es importante saber como se comporta la poblacion ante la enfermedad, como se comportan los individuos, como se comportan los grupos, como se comportan las familias, etc.

Otro de los objetivos es poder saber con antelacion su capacidad de producir la muerte en los individuos infectados. Esto es particularmente importante para mejorar la eficiencia en la medidas que se puede tomar para combatir la enfermedad.

La pandemia de COVID-19 fue la primer gran pandemia donde pudimos recolectar una gran cantidad de informacion. Es la primera vez en la historia de la humanidad donde se pudo rastrear y hacer seguimiento al curso de la enfermedad en millones de personas alrededor del mundo. Esta gran cantidad de datos nos permiten analizar y descubrir como se comportan las enfermedades cuando se propagan en una poblacion. Con esta gran cantidad de datos hoy podes predecir con mayor exactitud el comportamiento de la enfermedad. Intentaremos en este trabajo poder empezar a comprender uno de los aspectos mas criticos que es su mortalidad.

HIPOTESIS

El dataset elegido fue creado por el estado de Mexico, el cual contiene informacion de pacientes que fueron diagnosticados con COVID-19. El dataset contiene 21 columnas y 1,048,576 filas. Las columnas son las siguientes:

contenido del data set COVID

The dataset was provided by the Mexican government (link). This dataset contains an enormous number of anonymized patient-related information including pre-conditions. The raw dataset consists of 21 unique features and 1,048,576 unique patients. In the Boolean features, 1 means "yes" and 2 means "no". values as 97 and 99 are missing data.

- sex: 1 for female and 2 for male.
- age: of the patient.
- classification: covid test findings. Values 1-3 mean that the patient was diagnosed with covid in different
- degrees. 4 or higher means that the patient is not a carrier of covid or that the test is inconclusive.
- patient type: type of care the patient received in the unit. 1 for returned home and 2 for hospitalization.
- pneumonia: whether the patient already have air sacs inflammation or not.
- pregnancy: whether the patient is pregnant or not.
- diabetes: whether the patient has diabetes or not.
- copd: Indicates whether the patient has Chronic obstructive pulmonary disease or not.
- asthma: whether the patient has asthma or not.
- inmsupr: whether the patient is immunosuppressed or not.
- hypertension: whether the patient has hypertension or not.
- cardiovascular: whether the patient has heart or blood vessels related disease.
- renal chronic: whether the patient has chronic renal disease or not.
- other disease: whether the patient has other disease or not.
- obesity: whether the patient is obese or not.
- tobacco: whether the patient is a tobacco user.
- usmr: Indicates whether the patient treated medical units of the first, second or third level.
- medical unit: type of institution of the National Health System that provided the care.
- intubed: whether the patient was connected to the ventilator.
- icu: Indicates whether the patient had been admitted to an Intensive Care Unit.
- date died: If the patient died indicate the date of death, and 9999-99-99 otherwise.

Mediante la utilizacion de los distintos datos que nos presenta el dataset, se puede realizar una serie de hipotesis que nos permitiran analizar el comportamiento de los datos y asi poder realizar una prediccion de los mismos. Se tratara de encontrar mediante regresion LINEAL la relacion entre las diferentes enfermedades preexsitentes y los distintos grados de avance de la enfermedad, desde lo mas leve a los mas grave y finalmente la muerte. Tambien se analizara la relacion entre el sexo y la edad de los pacientes, asi como la relacion entre el tipo de paciente y el tipo de unidad medica en la que fue atendido. Esto ayudara a poder adaptar los tratamientos y cuidados de los pacientes de acuerdo a las necesidades de cada uno de ellos.

Preguntas a responder:

- ¿Cual es la relacion entre las diferentes enfermedades preexsitentes y los distintos grados de avance de la enfermedad?
- ¿Cual es la relacion entre el sexo y la edad de los pacientes?
- ¿Cual es la relacion entre el tipo de paciente y el tipo de unidad medica en la que fue atendido?

In [ ]:
```
df_covid = pd.read_csv('./Covid Data.csv')
```

In [ ]:
```
df_covid.head()
```

Out[ ]:

	USMER	MEDICAL_UNIT	SEX	PATIENT_TYPE	DATE_DIED	INTUBED	PNEUMONIA	AGE	PR
0	2	1	1	1	03/05/2020	97	1	65	
1	2	1	2	1	03/06/2020	97	1	72	
2	2	1	2	2	09/06/2020	1	2	55	
3	2	1	1	1	12/06/2020	97	2	53	
4	2	1	2	1	21/06/2020	97	2	68	

In [ ]:
```
df_covid.shape
```

Out[ ]:
```
(1048575, 21)
```

In [ ]:
```
df_covid.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1048575 entries, 0 to 1048574
Data columns (total 21 columns):
 #   Column                Non-Null Count    Dtype 
---  ------                --------------    ----- 
 0   USMER                 1048575 non-null  int64 
 1   MEDICAL_UNIT          1048575 non-null  int64 
 2   SEX                   1048575 non-null  int64 
 3   PATIENT_TYPE          1048575 non-null  int64 
 4   DATE_DIED             1048575 non-null  object
 5   INTUBED               1048575 non-null  int64 
 6   PNEUMONIA             1048575 non-null  int64 
 7   AGE                   1048575 non-null  int64 
 8   PREGNANT              1048575 non-null  int64 
 9   DIABETES              1048575 non-null  int64 
 10  COPD                  1048575 non-null  int64 
 11  ASTHMA                1048575 non-null  int64 
 12  INMSUPR               1048575 non-null  int64 
 13  HIPERTENSION          1048575 non-null  int64 
 14  OTHER_DISEASE         1048575 non-null  int64 
 15  CARDIOVASCULAR        1048575 non-null  int64 
 16  OBESITY               1048575 non-null  int64 
 17  RENAL_CHRONIC         1048575 non-null  int64 
 18  TOBACCO               1048575 non-null  int64 
 19  CLASIFFICATION_FINAL  1048575 non-null  int64 
 20  ICU                   1048575 non-null  int64 
dtypes: int64(20), object(1)
memory usage: 168.0+ MB
```

El dataset cuenta con las enfermedades y otros factores previos al contagio del COVID, así como el tipo de atención médica que recibió el paciente, si fue hospitalizado o no, si fue intubado o no, si falleció o no, y la fecha de fallecimiento. Con estos datos se puede analizar la relación entre las enfermedades previas y el resultado del COVID, así como la relación entre el tipo de atención médica y el resultado del COVID. Se intentara predecir

si el paciente falleció o no, y si fue intubado o no, con base en las enfermedades previas y el tipo de atención médica que recibió.

OBJETIVOS

El objetivo es poder comprender el comportamiento de la enfermedad para permitir la aplicacion de distintos tratamientos y la aplicacion de sistemas de triage para poder atender a los pacientes de acuerdo a sus necesidades. Reduciendo tiempos de espera, tiempos de tratamiento y tiempos de recuperacion. Esto se va a intentar mediante la prediccion de la mortalidad de los pacientes. Lo que va a permitir hacer mas eficiente y efectivo el uso de los recursos disponibles.

Graficas / EDA

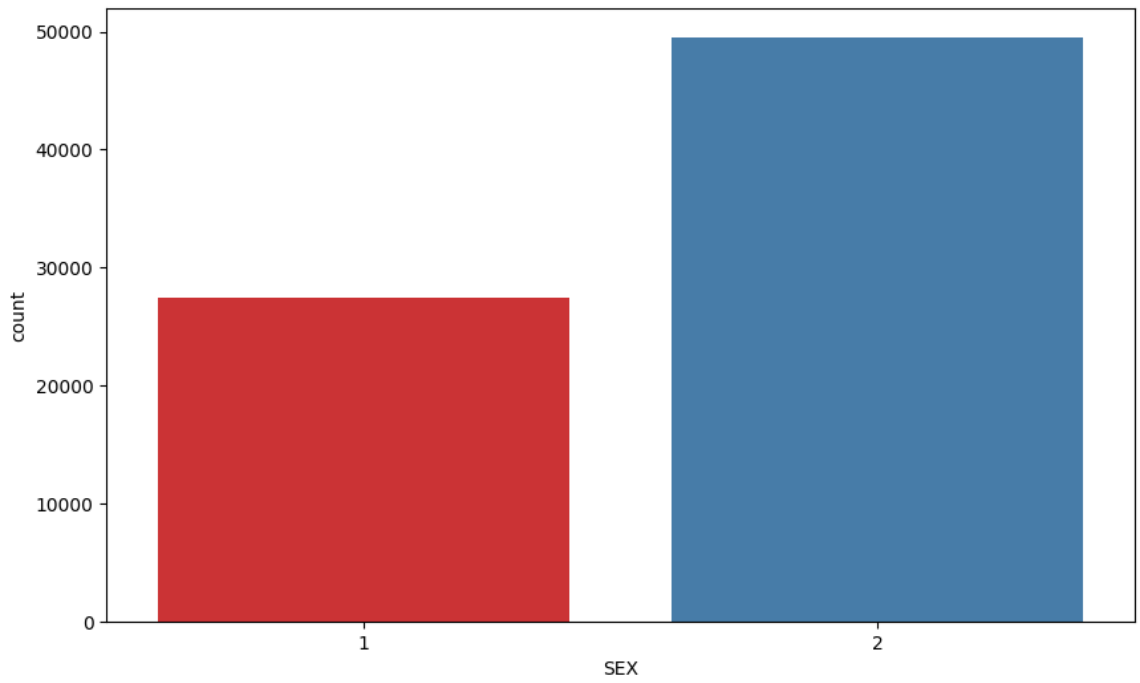
Vamos a empezar a analizar las distintas variables para poder tener una aproximacion de como se comportan los datos y poder hacer un analisis mas profundo mediante la hipotesis de que enfermedades previas al COVID pueden influir en el resultado del desarrollo de la enfermedad.

fallecidos por sexo

En la siguiente grafica se puede observar que el numero de fallecidos es mayor en el sexo masculino, esto puede deberse a que los hombres son mas propensos a tener enfermedades preexistentes que las mujeres, como diabetes, hipertension, etc.

In [ ]:
```
fig, ax = plt.subplots(figsize=(10, 6))
sns.countplot(x='SEX', data=df_covid[df_covid['DATE_DIED'] != "9999-99-99"], pal
```

Out[ ]:
```
<AxesSubplot: xlabel='SEX', ylabel='count'>
```

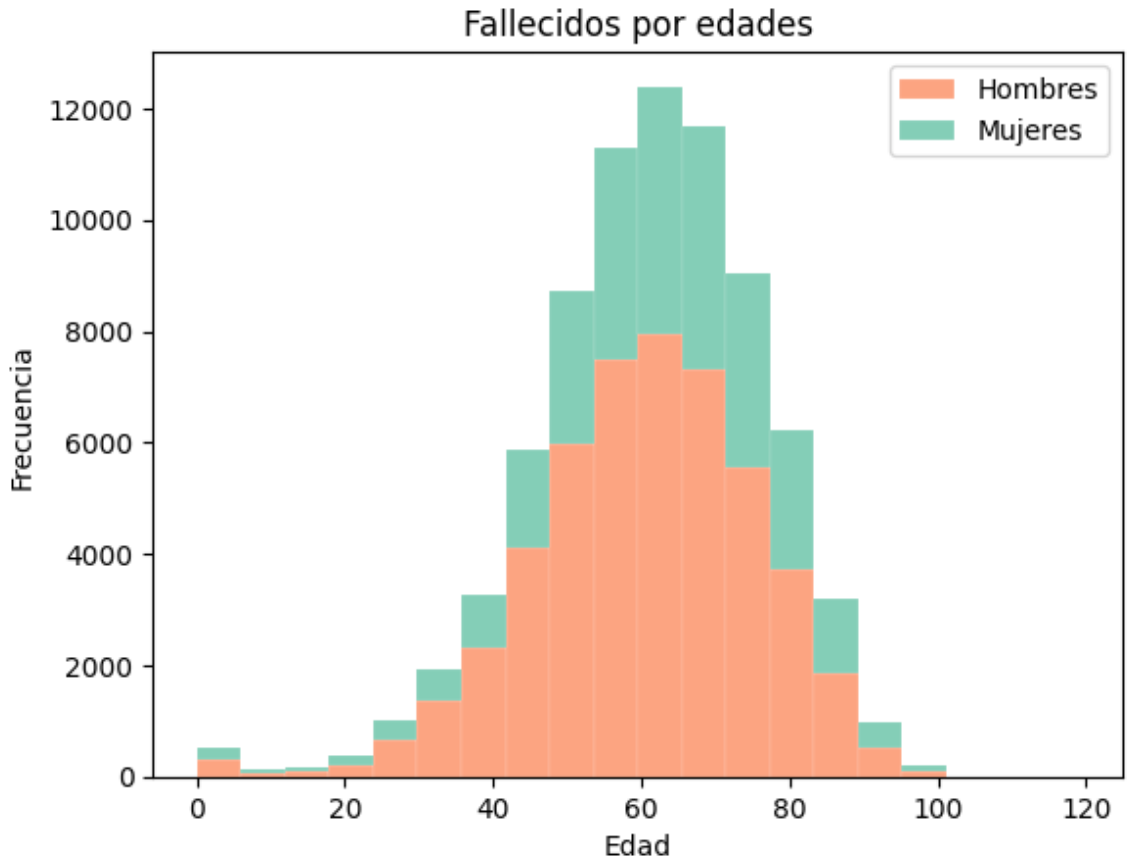


fallecidos por sexo y edad

En la grafica se observan los fallecidos por edad y por sexo. Se puede observar que la mayoria de los fallecidos son hombres, y que la mayoria de los fallecidos son mayores de 60 años.

In [ ]:
```
bar = sns.histplot(data=df_covid[df_covid['DATE_DIED'] != "9999-99-99"], x='AGE'
bar.set(xlabel='Edad', ylabel='Frecuencia', title='Fallecidos por edades')
bar.legend(['Hombres', 'Mujeres'])
```

Out[ ]:
```
<matplotlib.legend.Legend at 0x21d3d0259c0>
```

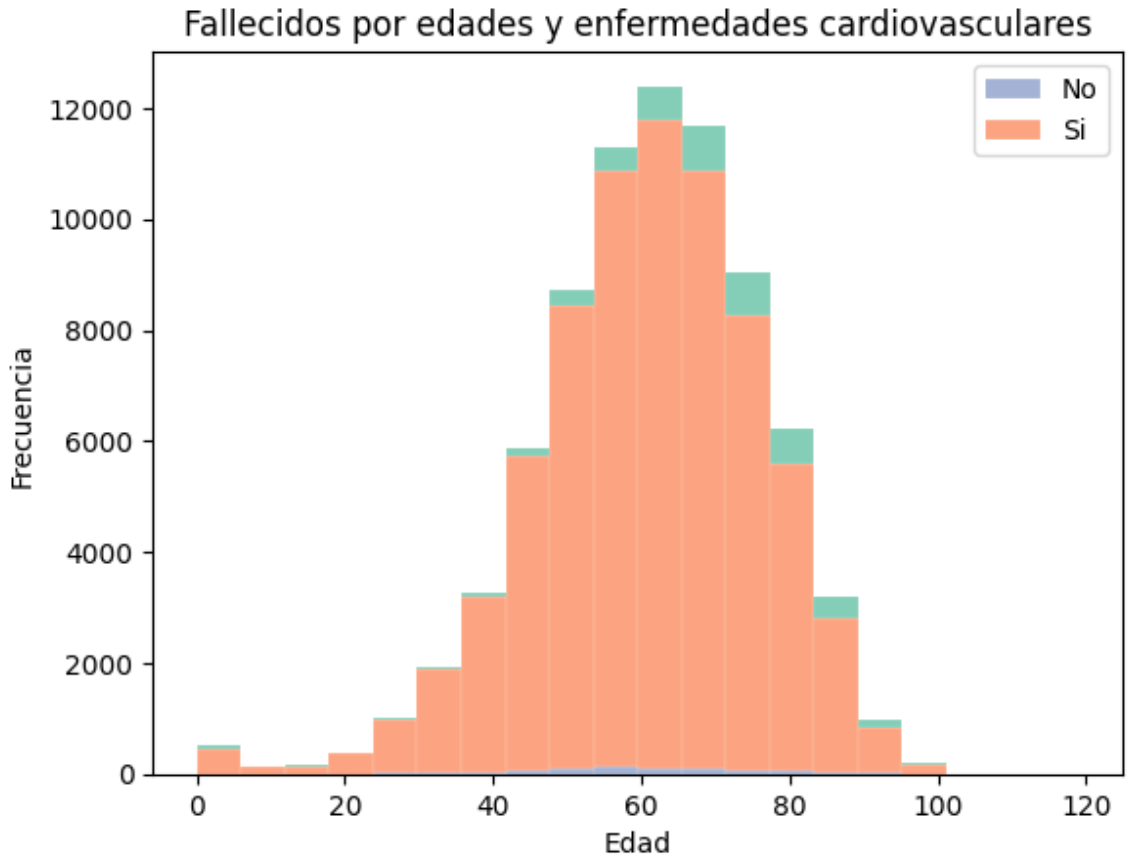


fallecidos por edad y enfermedades cardiovasculares

Analisamos los fallecidos por edad y con enfermedades cardiovasculares. Se puede observar que la mayoria de los fallecidos son mayores de 60 años y tienen enfermedades cardiovasculares.

In [ ]:
```
bar = sns.histplot(data=df_covid[df_covid['DATE_DIED'] != "9999-99-99"], x='AGE'
bar.set(xlabel='Edad', ylabel='Frecuencia', title='Fallecidos por edades y enfer
bar.legend(['No', 'Si'])
```

Out[ ]:
```
<matplotlib.legend.Legend at 0x21d3daf9bd0>
```



Del total con enfermedades cardiovasculares, sobrevivientes y fallecidos

Tomamos una enfermedad preexistente para empezar a comprender la incidencia de las mismas en el desarrollo de la enfermedad.

Comparamos el total de pacientes con enfermedades cardiovasculares y sin enfermedadades cardiovasculares. Se puede observar un porcentaje mayor de fallecidos en los pacientes con enfermedades cardiovasculares.

In [ ]:
```
total_cardio = df_covid[df_covid['CARDIOVASCULAR'] == 1]
print(total_cardio['CARDIOVASCULAR'].sum())
sin_cardio = df_covid[df_covid['CARDIOVASCULAR'] == 2]
```

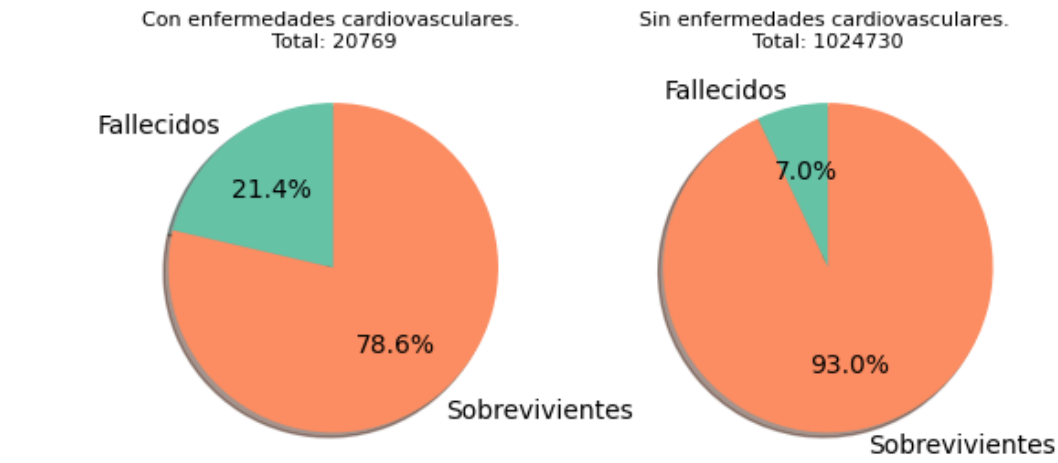
```
20769
```

In [ ]:
```
fallecidos_cardio = df_covid[(df_covid['CARDIOVASCULAR'] == 1) & (df_covid['DATE
sobrevivientes_cardio = df_covid[(df_covid['CARDIOVASCULAR'] == 1) & (df_covid['
fallecidos_sin_cardio = df_covid[(df_covid['CARDIOVASCULAR'] == 2) & (df_covid['
sobrevivientes_sin_cardio = df_covid[(df_covid['CARDIOVASCULAR'] == 2) & (df_cov
fig, ax = plt.subplots( ncols=2, nrows=1)
colores = sns.color_palette('Set2')

ax[0].pie([len(fallecidos_cardio), len(sobrevivientes_cardio)], labels=['Falleci
ax[0].set_title('Con enfermedades cardiovasculares. \nTotal: ' + str(len(total_c


ax[1].pie([len(fallecidos_sin_cardio), len(sobrevivientes_sin_cardio)], labels=[
ax[1].set_title('Sin enfermedades cardiovasculares. \nTotal: ' + str(len(sin_car
```

```
plt.suptitle('Porcentaje de fallecidos y sobrevivientes por enfermedades cardiov
plt.show()
```

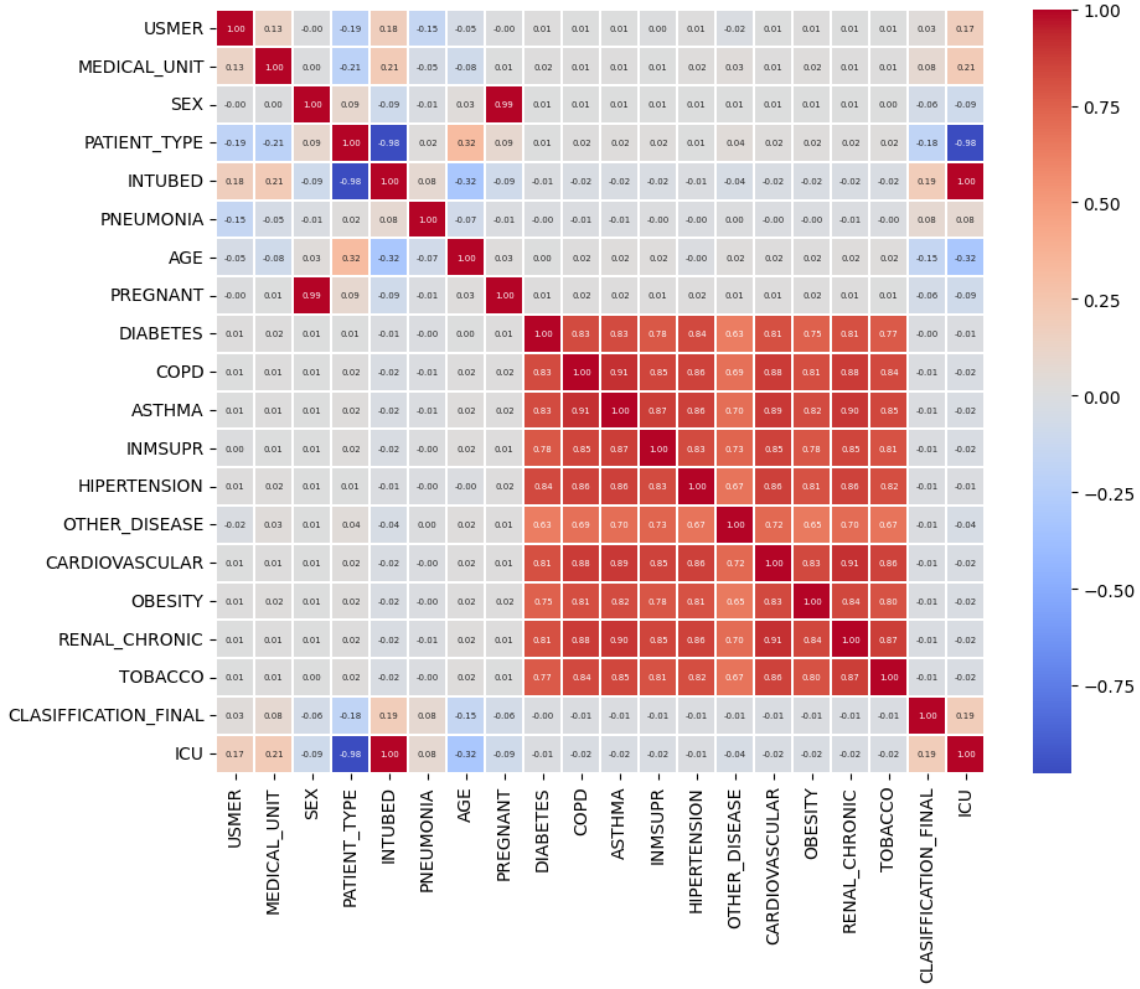


Mapa de calor para ver la correlacion entre las variables y sus posibles incidencias en el resultado del COVID

Generamos un mapa de calor para ver la correlacion entre las variables y sus posibles incidencias en el resultado del COVID. Se puede observar que las variables que tienen mayor correlacion con el resultado del COVID son las enfermedades cardiovasculares, diabetes, hipertension, y la edad.

In [ ]:
```
sns.heatmap(df_covid.corr(), annot=True, cmap='coolwarm', linewidths=0.2, annot_
fig=plt.gcf()
fig.set_size_inches(10,8)
plt.show()
```

```
C:\Users\ismael\AppData\Local\Temp\ipykernel_6660\2490466195.py:1: FutureWarnin
g: The default value of numeric_only in DataFrame.corr is deprecated. In a futu
re version, it will default to False. Select only valid columns or specify the
value of numeric_only to silence this warning.
  sns.heatmap(df_covid.corr(), annot=True, cmap='coolwarm', linewidths=0.2, ann
ot_kws={'size':5}, fmt='.2f')
```



Contexto Comercial

Los estados son la primer gran barrera de proteccion y cuidado ante la pandemia. Pero como todo su capacidad de respuesta depende mucho del conocmientos de la enfermedad. Los recursos con los que cuentan son finitos y cuando mas eficiente sea su uso mas vidas se salvaran y mas rapido se lograra el control de la enfermedad. Los estados fueron tambien los encargados de la recollecion de los datos y su publicacion. El presente trabajo permitira tener una mejor comprension de la enfermedad y permitir a los estados mejorar su capacidad de respuesta y adecuarse a ella.